FACTORS AFFECTING ONLINE NEWS POPULARITY

BIOST 557: Applied Statistics And Experimental Design

Team Members: Abhinav Duvvuri, Mihir Gathani, Navya Eedula, Sushma Vankayala, Swarali Desai

ABSTRACT

This study investigates the determinants of online news article popularity, focusing on the influence of multimedia content, publication day, title and content length, keyword usage, and data channels. Using a dataset of 39,644 articles from Mashable spanning two years, the research employed various statistical methods, including descriptive analysis, chi-square tests, logistic regression, Welch two-sample t-test, and analysis of variance (ANOVA), to explore the relationship between these factors and article shares, a proxy for popularity. The findings indicate significant differences in multimedia usage across categories, with images and videos significantly impacting article popularity. While the specific publication day did not significantly influence the popularity of articles, categorizing them into weekdays and weekends revealed a notable impact on their popularity. The study also found that articles with certain keywords and those published under specific data channels (e.g., "Lifestyle", "Social Media") tend to have higher share counts, highlighting the importance of content relevance and audience targeting. Moreover, the length of the article's title was positively associated with popularity, whereas the content length showed no significant effect. The research concludes that content creators should focus on incorporating relevant multimedia, choosing impactful keywords, publishing on the right day, and crafting engaging titles to enhance the popularity of online news articles. Future studies could further explore the intricate dynamics between content features and reader engagement, considering the rapidly evolving nature of online media consumption.

Table of Contents

1 Introduction

1.1 Background

With the expansion of the Internet, there has also been a growing interest in online news, which allows for the easy and fast spread of information. The internet provides a constant stream of news articles, making it difficult for audiences to keep up and make informed decisions about what news to consume. Hence, understanding the factors influencing popularity is valuable for authors, content providers, advertisers, and activists/politicians[1]. Predicting the popularity of online news has emerged as a recent research trend to better reach target audiences. Popularity is often measured by considering the number of interactions on the Web and social networks (e.g., number of shares, likes, and comments). By identifying what makes news popular, creators can focus on elements that resonate with the audience and increase engagement and potentially higher revenue. This could involve specific writing styles, headlines, content formats, or even publishing articles on certain days over others.

1.2 Project Goals

The project aims to investigate various factors influencing the popularity of articles. We aim to explore the relationship between image and video count and article popularity. Additionally, our goal is to examine the influence of data channels on article shares, to reveal if any significant effects on popularity exist. We seek to uncover patterns related to the day of publication (weekend, weekday, etc) to check if this correlates with higher shares. Furthermore, our project endeavors to analyze the impact of title length on popularity, recognizing variations across different categories. Lastly, we aim to understand the role of keywords in influencing article popularity, recognizing their differential impact across categories.

2 Dataset

2.1 Description

The selected dataset[2] comprises a total of 39,644 articles scraped from Mashable over two years, sourced from the UCI repository. These articles span from January 7th, 2013, to January 7th, 2015. The data contains a list of characteristics that describe different aspects of the article and that were considered relevant to influence the number of shares. The variables used in this study include shares, number of images, number of videos, type of data channel, day of publication, title length, content length, and keyword metrics. The data has already been preprocessed - categorical and boolean variables have been encoded, among other steps. Certain columns are calculated fields (ex: all the keyword-related fields).

2.2 Study Design

We assume that the articles present in this dataset are the population of all Mashable articles produced within the specified timeframe. We operate under the assumption that the calculated fields are reliable for our analysis. We confirm that there are no null values in the dataset. This dataset pertains to an observational study, wherein we've aimed to offer valuable insights into variable relationships. However, we acknowledge the limitations inherent in proving causation within observational studies.

For our analysis, we have used the number of shares as a measure of popularity. Additionally, in select analyses, we defined a threshold to distinguish popular from non-popular articles, considering the top 1% (those with over 30,000 shares) as popular. This user-defined column was utilized in analyses of logistic regression. All data visualizations, as well as statistical analyses, were conducted using statistical software (e.g., R, Python), and at a significance level set at $\alpha = 0.05$.

3 Statistical Methods

3.1 Q1: Does multimedia usage vary across categories, and does it impact article popularity?

- Sushma Vankayala

3.1.1 Data:

Four existing features were used for the analysis: number of images (*num_imgs*), number of videos (*num_videos*), category, and the popularity indicator (*is_popular*). For ease of analysis, new fields were created to indicate the presence or absence of media (*has_image*, *has_video*) and the standardized numbers of images and videos (*num_imgs_std*, *num_videos_std*).

3.1.2 Descriptive analysis

To understand the relationship between multimedia content and article categories, I utilized stacked bar graphs. These graphs visualized the proportion of articles within each category that contained at least one image or video (multimedia articles) compared to those without. By analyzing the graphs, it appeared that the proportion of articles with image content(Fig 3.1.1) and video content(Fig 3.1.2) varied across different categories.

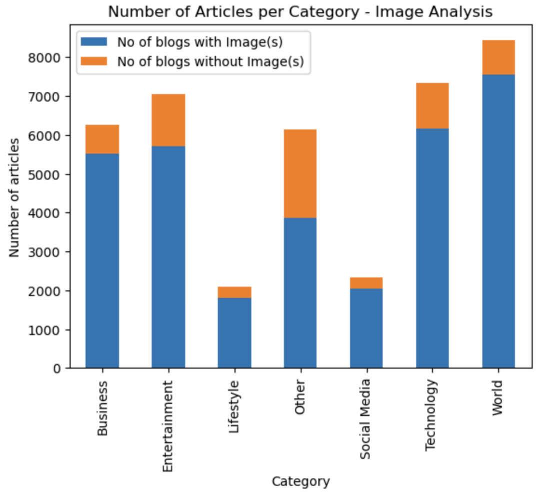


Fig 3.1.1 Number of articles per category. Color-coded by image presence.

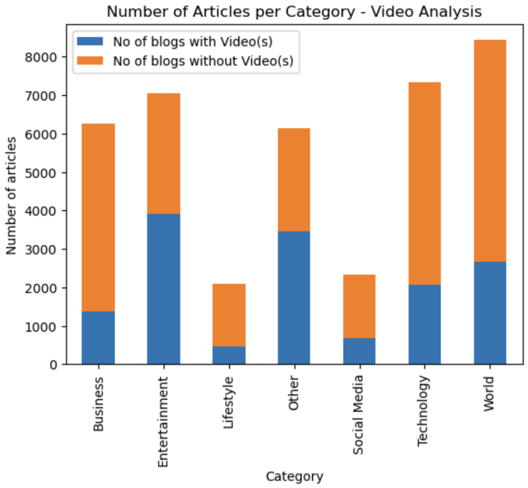


Fig 3.1.2 Number of articles per category. Color-coded by video presence.

3.1.3 Part A: Is the proportion of articles with images/videos different across categories?

3.1.3.1 Statistical Method:

Test Used: Chi-square for goodness-of-fit

Null Hypothesis (H_o): the proportion of articles with images is the same across all the categories.

Alternate Hypothesis (H_a): the proportion of articles with images is not the same across all the categories.

To facilitate the test, I calculated expected counts for articles with images in each category based on the overall observed proportion across all categories. All the assumptions for performing the Chi-square test were met -

- **Large sample size -** Data consisted of 39,644 articles, with each cell value greater than 10
- **Categorical Variables -** Both *category* and *has_image* variables are categorical
- **Mutually exclusive -** An article cannot belong to multiple categories.
- **Independence -** the presence of image(s) in one article doesn't influence the image presence in another.

3.1.3.2 Result:

- On performing the chi-square goodness-of-fit test, I obtained a test statistic of **375.61** and a p-value of **4.86e-78**. At a significance level of 0.05, I reject the null hypothesis and state that the proportions of articles(with images) among categories are not the same.
- A similar test was conducted for the proportions of articles with videos, and I obtained a test statistic of **2053.38** and a p-value of **0.0**. At a significance level of 0.05, I reject the null hypothesis and state that the proportions of articles(with videos) among categories are not the same.

3.1.4 Part B: Understanding the influence of multimedia content on the popularity of an article.

3.1.4.1 Statistical Method:

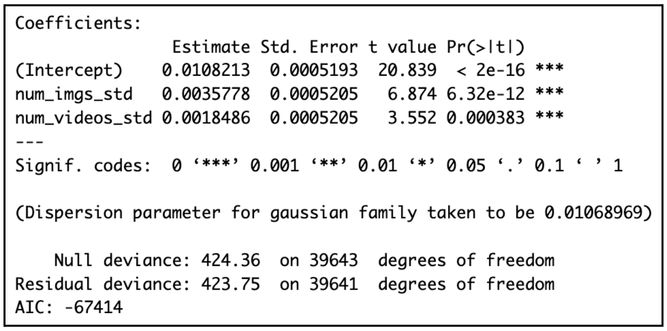
Since *is_popular* is a binary variable, I used a logistic regression model to understand the influence of multimedia content on article popularity. The below assumptions for logistic regression were met:

- Linearity with log odds (tested per explanatory variable)
- Absence of Multicollinearity (tested using correlation matrix)
- Independence of Observations
- Adequate Sample Size

First, I wanted to check if the standardized number of images and videos in an article influence the popularity of an article. The data was fit using the below model:

Model_1: *is_popular ~ num_imgs_std + num_videos_std*

The response variable is the log odds of an article being popular. The standardized number of images and videos in an article are considered as explanatory variables.

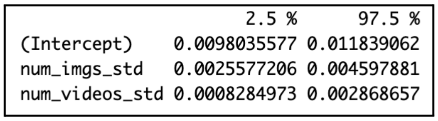


```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.0108213  0.0005193  20.839  < 2e-16 ***
num_imgs_std   0.0035778  0.0005205   6.874 6.32e-12 ***
num_videos_std 0.0018486  0.0005205   3.552 0.000383 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.01068969)

    Null deviance: 424.36  on 39643  degrees of freedom
Residual deviance: 423.75  on 39641  degrees of freedom
AIC: -67414
```

Fig 3.1.3: Summary of Model_1 denoting coefficients and deviance



```
                      2.5 %      97.5 %
(Intercept)    0.0098035577 0.011839062
num_imgs_std   0.0025577206 0.004597881
num_videos_std 0.0008284973 0.002868657
```

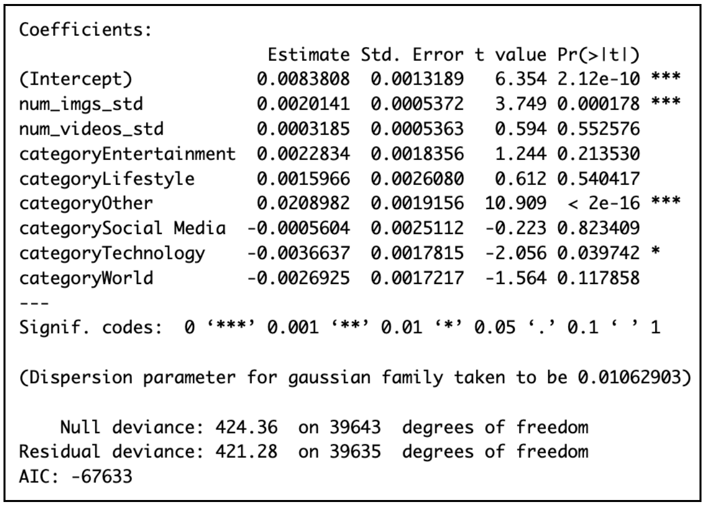
Fig 3.1.4: Confidence intervals of coefficients of Model_1

Understanding Coefficients: From Fig 3.1.3, we notice that both coefficients have extremely low p-values. This implies that there is a relationship between the standardized number of images and videos with an article being popular. Interpreting the coefficient of *num_imgs_std,* we can say that a unit increase in the *num_imgs_std* increases the log odds of an article being popular by 0.003 while adjusting for the number of videos(*num_videos_std)*. The coefficient of *num_videos_std* can be interpreted similarly.

Noticing that the residual deviance of Model_2 is close to the null deviance, I wanted to check if adding a *category* as an explanatory variable can improve the model.

Model_2: *is_popular ~ num_imgs_std + num_videos_std + category*

The response variable is the log odds of an article being popular. The standardized number of images, and videos in an article and the category of the article are considered explanatory variables. In this model, the "Business" category is considered as the reference value for the *category*.

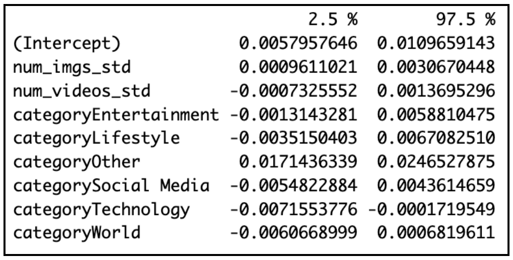


```
Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           0.0083808  0.0013189   6.354 2.12e-10 ***
num_imgs_std          0.0020141  0.0005372   3.749 0.000178 ***
num_videos_std        0.0003185  0.0005363   0.594 0.552576
categoryEntertainment  0.0022834  0.0018356   1.244 0.213530
categoryLifestyle     0.0015966  0.0026080   0.612 0.540417
categoryOther         0.0208982  0.0019156  10.909  < 2e-16 ***
categorySocial Media  -0.0005604  0.0025112  -0.223 0.823409
categoryTechnology    -0.0036637  0.0017815  -2.056 0.039742 *
categoryWorld         -0.0026925  0.0017217  -1.564 0.117858
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.01062903)

    Null deviance: 424.36  on 39643  degrees of freedom
Residual deviance: 421.28  on 39635  degrees of freedom
AIC: -67633
```

Fig 3.1.5: Summary of Model_2 denoting coefficients and deviance



```
                              2.5 %        97.5 %
(Intercept)            0.0057957646  0.0109659143
num_imgs_std           0.0009611021  0.0030670448
num_videos_std        -0.0007325552  0.0013695296
categoryEntertainment -0.0013143281  0.0058810475
categoryLifestyle     -0.0035150403  0.0067082510
categoryOther          0.0171436339  0.0246527875
categorySocial Media  -0.0054822884  0.0043614659
categoryTechnology    -0.0071553776 -0.0001719549
categoryWorld         -0.0060668999  0.0006819611
```

Fig 3.1.6: Confidence intervals of coefficients

Understanding Coefficients: As shown in Fig 3.1.5, Model_2 has a lower residual deviance in comparison to Model_1. In this model, we notice that the coefficient of *num_videos_std* is no longer significant. This means that when adjusted for *num_imgs_std* and *category*, the *num_videos_std* does not influence the popularity of an article. Interpreting the coefficient of *num_imgs_std*, we can say that a unit increase in the *num_imgs_std* increases the log odds of an article being popular by 0.002, while adjusting for the category(*category*) and number of videos(*num_videos_std*). We also note that at a significance level of 0.05, the coefficient of the “Other” category is different from 0 (p-value is less than 0.05, and the confidence intervals shown in Fig 3.1.6 do not contain 0). Interpreting the coefficient of *categoryOther,* we can say that in comparison to the "Business" category, articles in the "Other" category have slightly higher log odds of being popular by 0.02, after adjusting for the number of standardized images and videos.

3.1.4.2 Result:

From the above 2 models, we notice how the log odds of an article being popular were impacted by a combination of variables denoting the number of images, number of videos, and category. Between the 2 models, the model using all three variables seems to better explain the data (lower residual deviance). The coefficients of *num_imgs_std* and *categoryOther* are statistically significantly different from 0, which underscores that the odds of an article being popular are influenced by the number of images and categories. That said, we will need more domain knowledge to understand if their influence is of practical significance.

3.1.5 Limitations:

The response variable *is_popular* exhibits a significant class imbalance, with only a small portion of articles classified as popular. This might lead to models prioritizing the majority class (unpopular) during training. Our analysis can be generalized only if the sample data is representative of the entire population.

3.2 Q2: Do certain data channels produce less or more popular news articles online?

- Swarali Desai

3.2.1 Data

The dataset comprises multiple columns signifying the data channel of each article, marked by prefixes like Lifestyle, Entertainment, Business, Social Media, Technology, and World. Articles lacking an assigned channel have been categorized under 'Other' for this analysis. Each column utilizes binary indicators (0 or 1) to denote whether an article is associated with a specific channel. As expected, the 'shares' column serves as our metric for gauging the popularity of an article.

3.2.2 Statistical Methods

3.2.2.1 Descriptive

Analyzing the data reveals differences in the average popularity of articles among various channels.
Channels categorized as Other, Lifestyle, and Social Media exhibit higher average shares, implying that articles within these categories generally achieve greater popularity. Conversely, the World channel displays the lowest average shares, suggesting a lower popularity for its articles. Additionally, examining the variability in shares, it's evident that the Other and Business channels experience significant variability, while the Social Media channel demonstrates the least variability in shares. (Fig 3.2.1)

Sr. No.	Category	min	max	mean	std	var
1	Business	1	690400	3063.018536	15046.387626	2.263938e+08
2	Entertainment	47	210300	2970.487034	7858.133920	6.175027e+07
3	Lifestyle	28	208300	3682.123392	8885.017375	7.894353e+07
4	Other	4	843300	5945.189599	19392.998064	3.760884e+08
5	Social Media	5	122800	3629.383125	5524.167095	3.051642e+07
6	Technology	36	663600	3072.283283	9024.343803	8.143878e+07
7	World	35	284700	2287.734069	6089.669476	3.708407e+07

Table 3.2.1: Data description: Comparing the minimum, maximum, and mean shares across all the categories.

Due to a lack of detailed information about the exact category to which certain articles belong, they are classified under the "Other" category. This could include articles from newly emerging data sources or from sources that have not been categorized. Initial findings also indicate that the "Other" category shows considerable variation, implying that it could lead to inaccuracies in the broader analysis.

3.2.2.2 Inferential

For the given research question, I wish to compare the mean shares across all 7 populations of the data channels. Analysis of Variance is a way to provide a single test of comparing group means, which can be used here to identify if the means vary across the categories. ANOVA is suitable here because we're comparing the means of more than two groups (the different data channels) to see if at least one of them significantly differs from the others in terms of article popularity. First, I tested the assumptions that are to be satisfied for the ANOVA model, which is explained as follows:

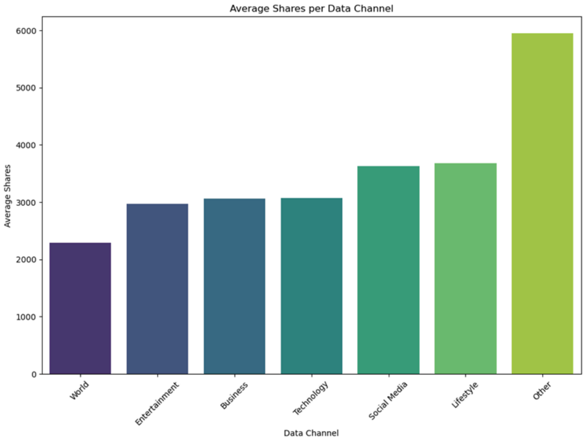
Independence: The model assumes that each observed share (as a percent of the total) in a particular category is independent of other categories, as one record has a unique category.

Variance: After observing unequal variances using Levene's test to check for homoscedasticity, we apply Welch's ANOVA to make sure our tests are robust against unequal variances

Normality/Sample Size: The sample sizes of the entire data and individual categories are sufficiently large (n=39644) and each, and the central limit theorem can be used to justify the normality of the sample.

Following are the null and alternative hypotheses for the ANOVA model:
Null Hypothesis (H0): The average number of shares is the same across all the data channels
Alternative Hypothesis (H1): The average number of shares is not the same across all the data channels

H0: μ_Entertainment = μ_Lifestyle = μ_Business = μ_Technology = μ_Other = μ_World = μ_SocialMedia
H1: At least one is different



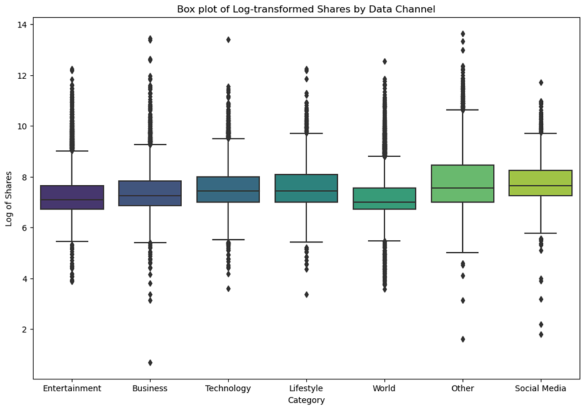
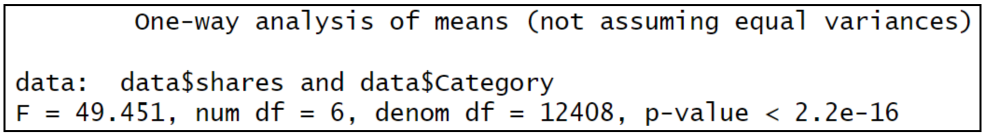


Fig 3.2.1: Average shares per Data Channel

Fig 3.2.2: Box plot for log shares per category

3.2.2.2.1 Results

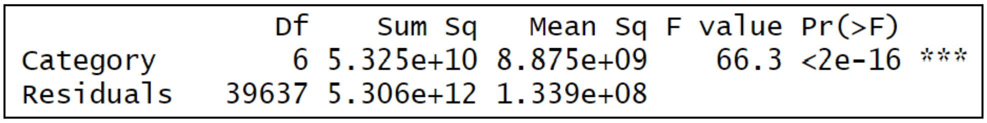
From the Global F-test results obtained using the ANOVA model (Fig 3.2.3 & Fig 3.2.4) with and without the equal variance assumption, we can conclude that the p-value for the F-test is very low (p-value < 2.2e-16) indicating that the results are statistically significant if we consider the significance level to be 5%. Hence, we can reject the null hypothesis that the average number of shares is the same across all the data channels.



```
        One-way analysis of means (not assuming equal variances)

data:  data$shares and data$Category
F = 49.451, num df = 6, denom df = 12408, p-value < 2.2e-16
```

Fig 3.2.3: Welch’s ANOVA with not equal variances



```
               Df    Sum Sq   Mean Sq F value Pr(>F)    
Category        6 5.325e+10 8.875e+09    66.3 <2e-16 ***
Residuals   39637 5.306e+12 1.339e+08                   
```

Fig 3.2.4: ANOVA with equal variances

3.2.2.3 Post-Hoc Analysis

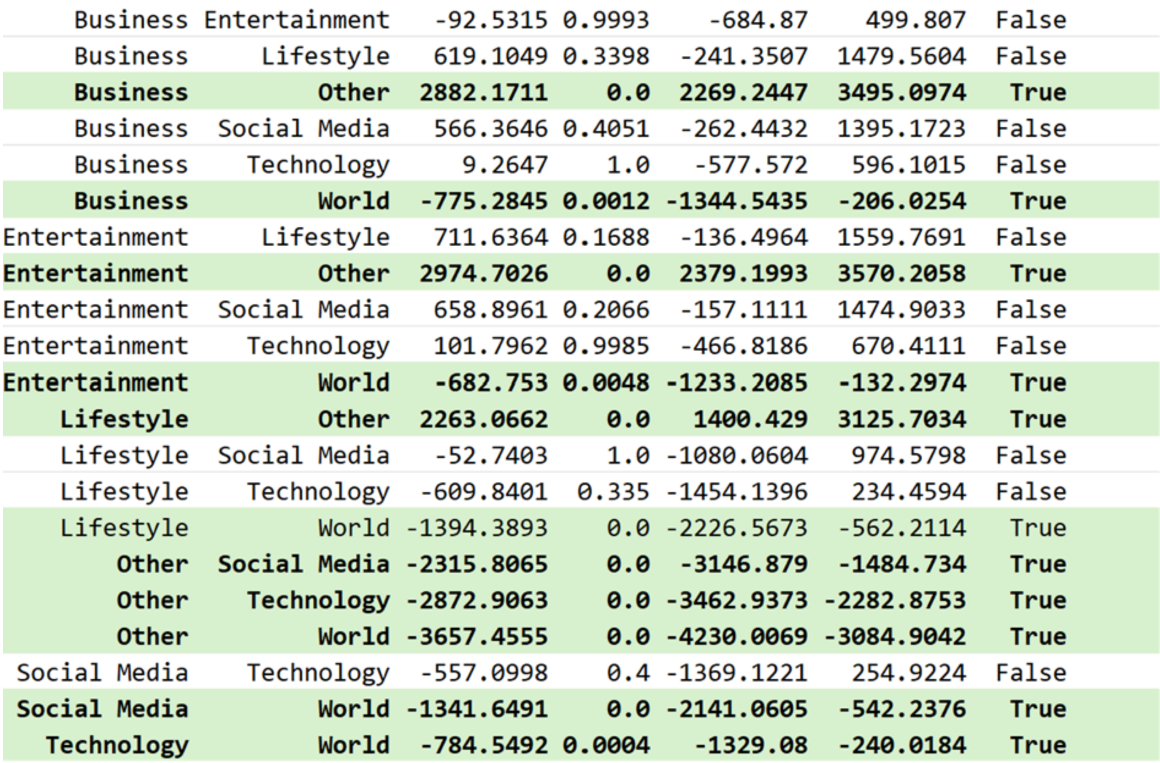
From the above results, we understand that the population means of the different categories are not the same. To understand which category is the most significant we need to perform some pairwise comparison. The problem with direct pairwise comparison is that it can inflate our type-I error. Thus, we need to perform adjustments to control the Family-wise errors. For this analysis, I performed the Tukey's Honestly Significant Difference (Tukey's HSD) test. This test compares all possible pairs of means while controlling for the Type I error rate across multiple comparisons. This test is conservative and has an assumption of equal variances within the group, an assumption considered valid for this analysis.

The output table below lists all pairs of groups, their mean differences, the confidence interval of the difference, and whether the difference is statistically significant (Reject column). A True in the Reject column indicates a statistically significant difference between the pair of groups at the specified alpha level, while False indicates no significant difference.

3.2.2.3.1 Results

The "Other" and "World" categories show significant differences when compared to most other categories, suggesting unique characteristics or factors influencing the number of shares in these categories that are not present in the comparison groups (Table 3.2.2). Articles categorized under "Other," "Lifestyle," and "Social Media" generally achieve greater popularity, while the "World" channel exhibits the lowest average shares, indicating less popularity.

```
                Multiple Comparison of Means - Tukey HSD, FWER=0.05
===========================================================================
     group1          group2        mean_diff    p-adj      lower       upper      reject
-------------------------------------------------------------------------------------------
```



```
    Business Entertainment   -92.5315 0.9993    -684.87    499.807  False
    Business     Lifestyle   619.1049 0.3398  -241.3507  1479.5604  False
    Business         Other  2882.1711    0.0  2269.2447  3495.0974   True
    Business  Social Media   566.3646 0.4051  -262.4432  1395.1723  False
    Business    Technology     9.2647    1.0   -577.572   596.1015  False
    Business         World  -775.2845 0.0012 -1344.5435  -206.0254   True
Entertainment    Lifestyle   711.6364 0.1688  -136.4964  1559.7691  False
Entertainment        Other  2974.7026    0.0  2379.1993  3570.2058   True
Entertainment Social Media   658.8961 0.2066  -157.1111  1474.9033  False
Entertainment   Technology   101.7962 0.9985  -466.8186   670.4111  False
Entertainment        World   -682.753 0.0048 -1233.2085  -132.2974   True
    Lifestyle        Other  2263.0662    0.0   1400.429  3125.7034   True
    Lifestyle Social Media   -52.7403    1.0 -1080.0604   974.5798  False
    Lifestyle   Technology  -609.8401  0.335 -1454.1396   234.4594  False
    Lifestyle        World -1394.3893    0.0 -2226.5673  -562.2114   True
        Other Social Media -2315.8065    0.0  -3146.879  -1484.734   True
        Other   Technology -2872.9063    0.0 -3462.9373 -2282.8753   True
        Other        World -3657.4555    0.0 -4230.0069 -3084.9042   True
 Social Media   Technology  -557.0998    0.4 -1369.1221   254.9224  False
 Social Media        World -1341.6491    0.0 -2141.0605  -542.2376   True
   Technology        World  -784.5492 0.0004   -1329.08  -240.0184   True
```

Table 3.2.2: Tukey's HSD test results

3.2.3 Conclusion

It can be concluded that the data channel significantly influences the number of shares and overall popularity. A limitation of this analysis is the potential need for a less conservative pairwise comparison test. As a next step, we should delve deeper into the "Other" category and get data to yield more actionable and relevant outcomes.

3.3 Q3: How do the publication day and channel affect the popularity of a news article?

- Mihir Gathani

3.3.1 Introduction and Data Choices

Aim: The objective of this analysis is to determine whether news articles published on certain days of the week are more likely to garner higher shares. Additionally, I aim to identify if particular news channels exhibit better performance on these high-sharing days

Data Choices: The dataset used here is a subset of the mashable dataset and includes multiple columns indicating the total shares for each news article, the day it was published, and the data channel it belongs to. The "weekday_is_dayofweek" column consists of binary indicators such as "weekday_is_monday," "weekday_is_tuesday," and so forth, representing the specific day of publication. These indicators have been merged into a single "day" column for enhanced usability, denoting the exact day of publication. Similarly, the channels category, represented by columns such as "data_channel_is_channelname," (Ex: data_channel_is business or data_channel_is_world) has been consolidated into a new column labeled "Channel." In instances where an article lacks a pre-classified channel, it has been categorized as "other."

3.3.2 Descriptive analysis

To investigate the correlation between article popularity (quantified by total shares) and its publication day, I conducted exploratory analysis employing both bar and box charts with mean markers, depicted in Figures 1 and 2, respectively. The findings revealed that while weekdays saw a higher volume of published articles, those published on weekends exhibited a notably higher average share count. This observation served as the focal point for subsequent analysis.

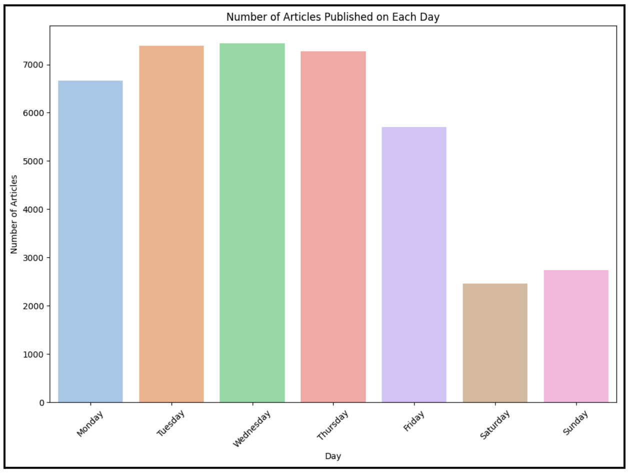


Figure 3.3.1:Number of articles published on each day

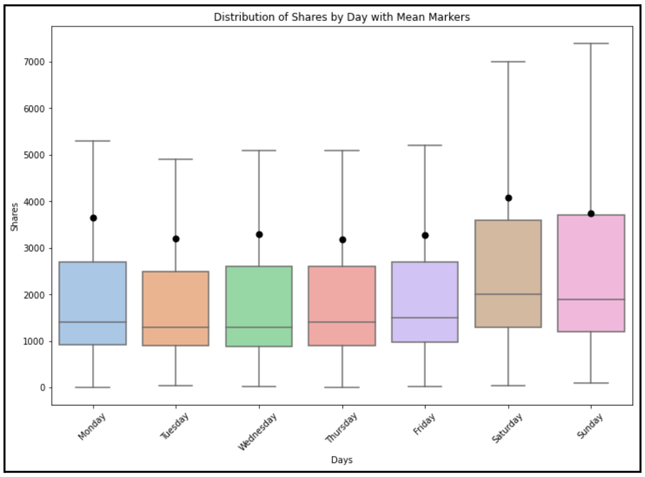


Figure 3.3.2: Distribution of shares by day with mean markers

3.3.3 Does news article popularity vary significantly among different days of the week?

3.3.3.1 Statistical Method:

For this research question, I wanted to compare the mean shares across all seven days to see if there was a significant difference. To do this, I utilized Analysis of Variance (ANOVA). I believe this is an appropriate method as it allows one to compare the means of three or more groups to determine whether they have statistically significant differences. In this case, the groups are the weekdays (Monday - Sunday) and we want to determine if there are significant differences in the average number of shares across these weekdays. Here, all assumptions for ANOVA are met as follows:

Independence: The model assumes that the number of shares observed for an article published on a particular day is independent of the number observed for articles published on other days.
Variance: I confirm homoscedasticity using Levene's test, allowing us to use ANOVA.
Normality: The sample sizes for each day are greater than 30, hence the central limit theorem can be used to justify the normality assumption.

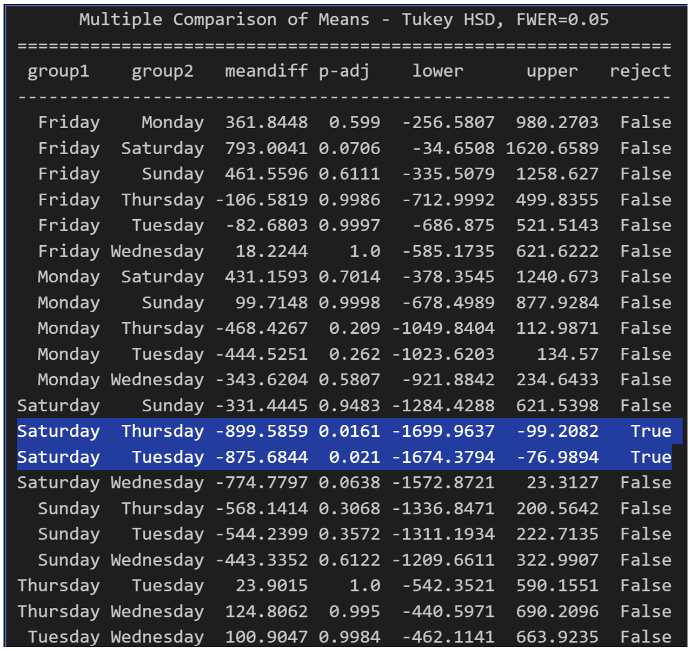
The null and alternative hypotheses would be as follows:
Null Hypothesis (H_0): There is no significant difference in the popularity of articles across days.
Alternative Hypothesis (H_a): There is a significant difference in the popularity of articles across days.

3.3.3.2 Results:

Using the ANOVA model and 0.05 significance level, I observed an F-statistic of 3.2705 and a p-value of 0.0032 indicating that there is a significant difference in the popularity of news articles across different publication days. Hence we can reject the null hypothesis.

3.3.3.3 Post-Hoc Analysis:

From the above results, we understand that the popularity of news articles across publication days is not the same. Thus we need to perform some pairwise comparison to understand which day is the most significant. The problem with direct pairwise comparison is it can inflate our type-I error. Thus, we can perform adjustments to control the Family-wise errors while computing the significance difference. For this analysis, I performed the Tukey's Honestly Significant Difference (Tukey's HSD) test. This test compares all possible pairs of means while controlling for the Type I error rate across multiple comparisons. I can use this based on the previous assumptions from ANOVA, and since it is robust to unequal sample size. This analysis concluded that Saturday performs better than Thursday and Tuesday. We can also observe the mean difference in the confidence interval for the same in Figure 3.



```
      Multiple Comparison of Means - Tukey HSD, FWER=0.05
================================================================
 group1    group2   meandiff p-adj    lower      upper   reject
----------------------------------------------------------------
  Friday    Monday  361.8448  0.599  -256.5807  980.2703  False
  Friday  Saturday  793.0041 0.0706   -34.6508 1620.6589  False
  Friday    Sunday  461.5596 0.6111  -335.5079  1258.627  False
  Friday  Thursday -106.5819 0.9986  -712.9992  499.8355  False
  Friday   Tuesday  -82.6803 0.9997   -686.875  521.5143  False
  Friday Wednesday   18.2244    1.0  -585.1735  621.6222  False
  Monday  Saturday  431.1593 0.7014  -378.3545  1240.673  False
  Monday    Sunday   99.7148 0.9998  -678.4989  877.9284  False
  Monday  Thursday -468.4267  0.209 -1049.8404  112.9871  False
  Monday   Tuesday -444.5251  0.262 -1023.6203    134.57  False
  Monday Wednesday -343.6204 0.5807  -921.8842  234.6433  False
Saturday    Sunday -331.4445 0.9483 -1284.4288  621.5398  False
Saturday  Thursday -899.5859 0.0161 -1699.9637  -99.2082   True
Saturday   Tuesday -875.6844  0.021 -1674.3794  -76.9894   True
Saturday Wednesday -774.7797 0.0638 -1572.8721   23.3127  False
  Sunday  Thursday -568.1414 0.3068 -1336.8471  200.5642  False
  Sunday   Tuesday -544.2399 0.3572 -1311.1934  222.7135  False
  Sunday Wednesday -443.3352 0.6122 -1209.6611  322.9907  False
Thursday   Tuesday   23.9015    1.0  -542.3521  590.1551  False
Thursday Wednesday  124.8062  0.995  -440.5971  690.2096  False
 Tuesday Wednesday  100.9047 0.9984  -462.1141  663.9235  False
```

Figure 3.3.3: Tukey's HSD Result

3.3.4 Does news article popularity vary significantly between weekdays and weekends?

3.3.4.1 Statistical Method:

From the previous analysis, we concluded that news articles published on Saturday have higher popularity in comparison with those published on Tuesday and Thursday. This prompted me to explore further to ascertain if weekends generally outperform weekdays. To address this query, I employed the Welch Two-Sample T-test. This method is well-suited for comparing two independent samples with potentially unequal variances and sample sizes.

Here, all assumptions for the Welch Two-Sample T-test are met as follows:
Independence: The model assumes that the number of shares observed for articles published on weekdays is independent of the number observed for articles published on weekends.
Variance: Welch's Two-Sample T-test is robust to unequal variances.
Normality: Although the Welch test is robust to normality, the sample sizes for both groups are greater than 30, hence the central limit theorem can be used to justify the normality assumption.

The null and alternative hypotheses would be as follows:
Null Hypothesis (H_0): There is no significant difference in the popularity of articles across weekdays and weekends.
Alternative Hypothesis (H_a): There is a significant difference in the popularity of articles across weekdays and weekends.

3.3.4.2 Results:

With the Welch Two-Sample T-test and a significance level of 0.05, I obtained a t-statistic of -3.5994 and a p-value of 0.0003, suggesting a significant difference in the average number of shares between news articles published on weekdays and those published on weekends. The mean difference indicates that news articles published on weekends perform better by an average of 584.54 shares compared to those published on weekdays. The 95% confidence interval for this mean difference is (266.19, 902.89). Hence, we can reject the null hypothesis and conclude that weekends outperform weekdays in garnering shares for news articles.

3.3.5 Are there specific news channels that outperform the rest on the weekends?

3.3.5.1 Statistical Method:

After concluding from the previous analysis that news articles published on weekends exhibit higher popularity, I sought to investigate if specific channels perform better on Saturday and Sunday. To do this, I employed Welch's Analysis of Variance (ANOVA), which is suitable for comparing means across multiple groups while accommodating unequal variances. In this instance, we compare the average number of shares across different data channels (e.g., business, world) on Saturday and Sunday.

Here, assumptions for Welch ANOVA are met as follows:
Independence: The model assumes that the number of shares for articles published within a channel is independent of those published in other channels.
Variance: Since Levene's test for homoscedasticity fails, we use Welch's ANOVA to ensure robustness
Normality: The sample sizes for each channel are greater than 30, hence the central limit theorem can be used to justify the normality assumption.

The null and alternative hypotheses would be as follows:

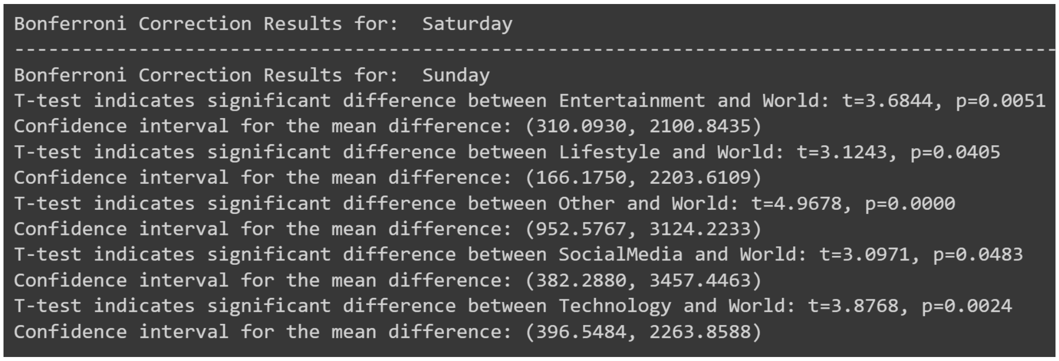
Null Hypothesis (H_0): There is no significant difference in the popularity of articles across different channels on a particular day (here, either Saturday or Sunday).
Alternative Hypothesis (H_a): There is a significant difference in the popularity of articles across different channels on a particular day (here, either Saturday or Sunday).

3.3.5.1 Results:

Utilizing the Welch ANOVA model with a significance level of 0.05, I obtained an F-statistic of 3.4067 and a p-value of 0.0025 for Saturday. Similarly, for Sunday, the F-statistic was 6.6263 with a p-value of 7.26e-07. These results indicate a significant difference in the popularity of news articles across different channels on both Saturday and Sunday. Therefore, allowing us to reject the null hypothesis.

3.3.5.1 Post-Hoc Analysis:

From the above results, it's evident that the popularity of news articles across various channels differs significantly on Saturday and Sunday. To delve deeper into identifying which channel holds the most significance on each day, pairwise comparisons are necessary. For this analysis, I performed pairwise t-tests with Bonferroni correction. This test compares all possible pairs of means while controlling for the Type I error rate across multiple comparisons. The analysis revealed that news articles from the "world" channel exhibited significantly lower performance compared to other channels, excluding business, on Sunday, as illustrated in Figure 4. However, the analysis did not identify any specific channel that outperformed others on Saturday. This result is attributed to the conservative nature of Bonferroni correction, which aims to control the overall Type I error rate across multiple comparisons.



```
Bonferroni Correction Results for:  Saturday
--------------------------------------------------------------------------------------------
Bonferroni Correction Results for:  Sunday
T-test indicates significant difference between Entertainment and World: t=3.6844, p=0.0051
Confidence interval for the mean difference: (310.0930, 2100.8435)
T-test indicates significant difference between Lifestyle and World: t=3.1243, p=0.0405
Confidence interval for the mean difference: (166.1750, 2203.6109)
T-test indicates significant difference between Other and World: t=4.9678, p=0.0000
Confidence interval for the mean difference: (952.5767, 3124.2233)
T-test indicates significant difference between SocialMedia and World: t=3.0971, p=0.0483
Confidence interval for the mean difference: (382.2880, 3457.4463)
T-test indicates significant difference between Technology and World: t=3.8768, p=0.0024
Confidence interval for the mean difference: (396.5484, 2263.8588)
```

Figure 3.3.4: Two-sample T-test with Bonferroni Correction Results

3.4 Q4: Do keywords affect the popularity of an article?

- Abhinav Duvvuri

3.4.1 Introduction and Data Choices

Knowing that Google uses keywords in its search engine optimization algorithm, I was interested to see if associations exist between keywords and the popularity of an article; even on the online news platform Mashable. Since I aimed to check for the popularity of an article, I will be using the popularity indicator field(is_popular) as described in the introduction.

The dataset contained a total of nine fields that provide information about keywords in an article. All keywords in an article are ranked based on their popularity and the average, maximum, and minimum number of shares(from historical data) of the best, worst, and average keywords in every article is available in the dataset. Out of these, only three fields made logical sense to use in this analysis. These are

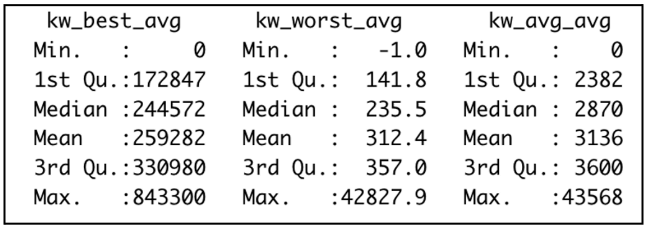
a. Avg. number of shares of the best keyword in an article, based on historical data*(kw_best_avg)*

b. Avg. number of shares of the worst keyword in an article, based on historical data*(kw_worst_avg)*
c. Avg. number of shares of the average keyword based on historical data *(kw_avg_avg)*

Apart from these three I also used another field num_keywords which describes the number of keywords in an article. The other six fields numerically encode the maximum and minimum number of shares for a particular keyword (based on historical data) which are extremes and do not represent the bulk of the data. Also, by design, since these attributes check the maximum and minimum number of shares for best/average/worst keywords in an article, they will have high multicollinearity with the above-mentioned fields. To avoid the risk of overfitting the data, I narrowed down my analysis to use only the four variables mentioned above.

3.4.2 Descriptive Analysis

The *num_keyword* field takes values between 1 and 10. With a mean of around 7 and a standard deviation of 1.9. There are no particularly big outliers. Figure 3.4.1 shows the distribution of the number of average shares (calculated using historical data) of the best, average, and worst keywords. The distribution of both *kw_best_avg* and *kw_avg_avg* seems logical, *kw_best_avg* is skewed towards the left and *kw_avg_avg* is skewed towards the right. I discovered that *kw_worst_avg* had negative values, which wouldn't be representative of real data. To ensure consistency, I removed the 600 affected rows (1.5% of the data) which had negative values in *kw_worst_avg*. Another thing to note is that the target variable *is_popular* has a class imbalance, with only 429 rows classified as popular compared to the rest being unpopular. This is in line with how very few articles become popular in real life.



```
  kw_best_avg       kw_worst_avg        kw_avg_avg
 Min.   :     0   Min.   :   -1.0   Min.   :    0
 1st Qu.:172847   1st Qu.:  141.8   1st Qu.: 2382
 Median :244572   Median :  235.5   Median : 2870
 Mean   :259282   Mean   :  312.4   Mean   : 3136
 3rd Qu.:330980   3rd Qu.:  357.0   3rd Qu.: 3600
 Max.   :843300   Max.   :42827.9   Max.   :43568
```

Figure 3.4.1: Descriptive statistics of kw_best_avg, kw_worst_avg, and kw_avg_avg

Since *is_popular* is a binary field, I chose to apply **logistic regression** to understand the influence of keywords on popularity. Before applying Logistic regression, I wanted to plot the distribution of explanatory variables for popular and non-popular articles and understand the potential differences in the mean and spread of these explanatory variables. Figure 3.4.2 shows the distribution of the explanatory variables for popular and non-popular articles. From these boxplots alone, on-average popular articles have a higher mean and median for *kw_avg_avg* and *kw_best_avg* attributes. A very small increase is seen in the case of *kw_worst_avg*. For the number of keywords, there is no visible difference in mean or median.

3.4.3 Inferential Analysis:

3.4.3.1 Assumptions of Logistic Regression

To proceed further, I tested for the assumptions of logistic regression. They are listed below:

1. **Binary dependent variable:** This is satisfied as we have two distinct categories popular and non-popular in our dependent variable.

2. **Independence of observations:** Since each article is independent of the other, we might assume independence. However, in cases where there are links to other articles inside an article, there is a chance that if the former article becomes popular, there is a higher chance that the linked article will also become popular as it is more accessible. However, for the analysis, I assumed that independence is valid.

3. **Adequate sample size:** Since the dataset is very large, this assumption is also met.
4. **Absence of multicollinearity:** We have already discussed this. The four explanatory variables have been chosen in such a way that the multicollinearity will be as low as possible.

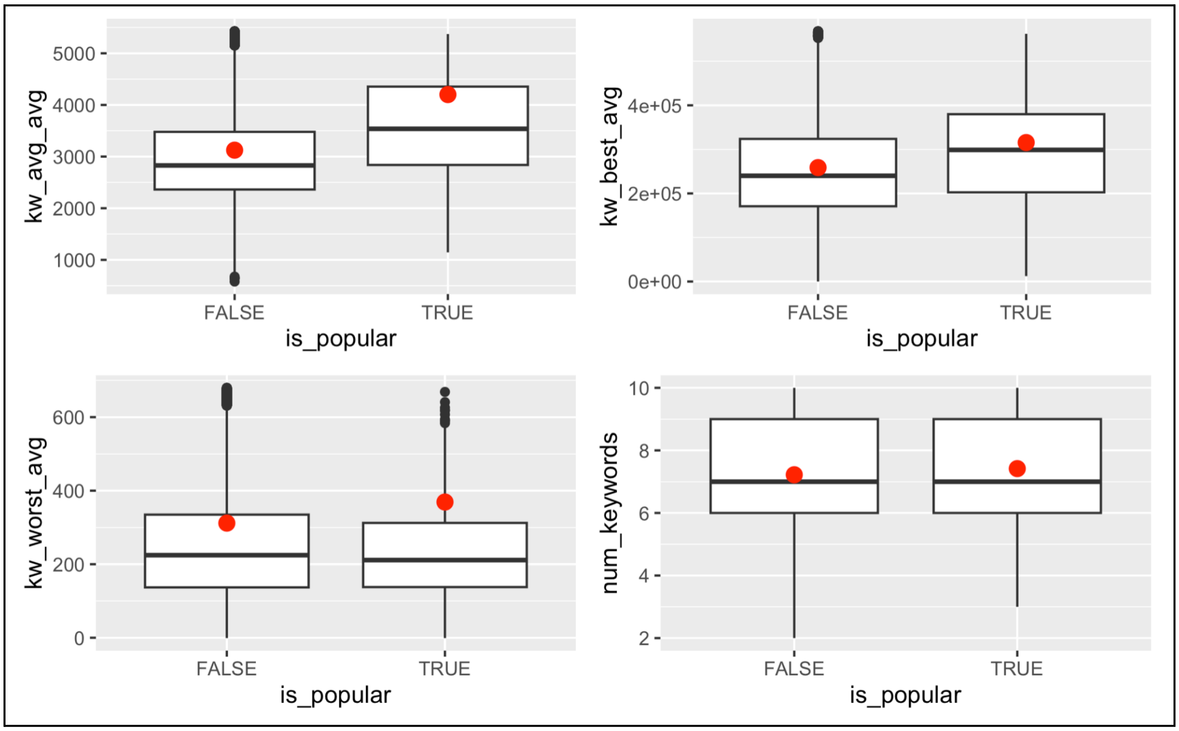


Figure 3.4.2: The distribution of the explanatory variables num_keyword, kw_avg_avg, kw_best_avg, and kw_worst_avg for both popular and non-popular article groups. The red dot shows the mean. (Outliers removed)

3.4.3.2 Models

1. **Logistic Model 1:** *is_popular ~ num_keyword + kw_avg_avg + kw_best_avg + kw_worst_avg*

   I ran logistic regression to understand the effect of the explanatory variables on the logit probability of an article being popular without interaction.

2. **Logistic Model 2:** *is_popular ~ kw_avg_avg*kw_best_avg*kw_worst_avg + num_keywords*

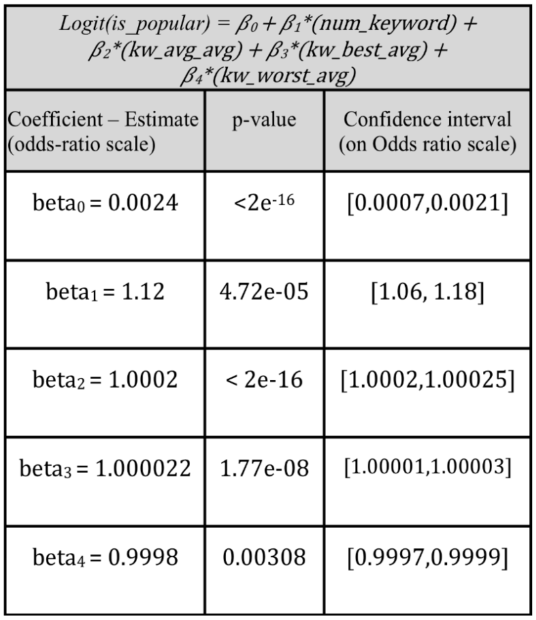
   I wanted to check if the interaction of these terms would yield a better model. For this, I fit the model by considering interactions among *kw_best_avg*, *kw_worst_avg,* and *kw_avg_avg*. I did not add *num_keyword* to the interaction because logically, there is no relation between the number of keywords in an article and a keyword being best, avg, or worst.

3. **Likelihood ratio test:** to understand if the interaction term made the model better, I ran an LRT (Likelihood ratio) test with the full model including the interaction variable, and the reduced model without the interaction variables.

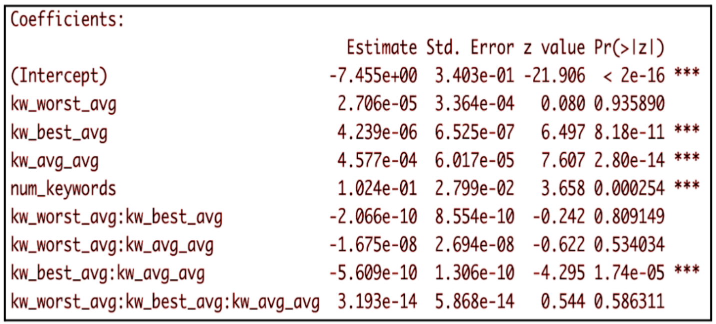
3.4.3.3 Results

1. Logistic Model 1: The deviance, AIC, and BIC of this model are listed in Table 3.4.2. The residual deviance decreased considerably from the null model. The p-values observed for all four variables had statistically significant beta values at the 0.05 significance level. For the point estimates and confidence intervals, I used the odds ratio scale as it is more interpretable. Table 3.4.1 shows the confidence intervals, point estimates, and p-values obtained. The model estimates an Odds Ratio of 1.0002 per 1 share difference in kw_avg_avg i.e. **the estimated odds of being popular are 0.0002% higher per 1 share difference adjusting for all other variables**. The other variables can be interpreted similarly.

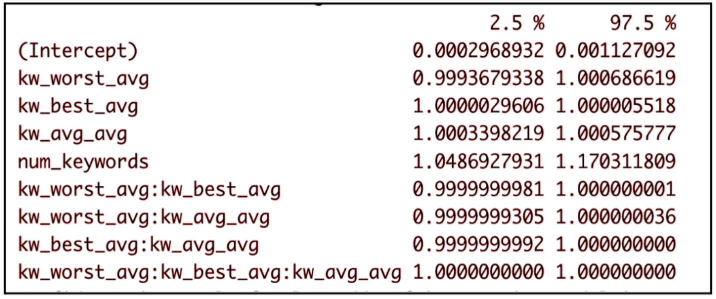
2. Logistic Model 2: The coefficients of the new model containing interactions are shown in Figure 3.4.3. While *kw_best_avg*, *kw_avg_avg*, *num_keyword,* and the interaction between *kw_best_avg* and *kw_avg_avg* had a **statistically significant p-value** at the 0.05 significance level, all other results were not statistically significant. This is logical as the best keyword and the average keyword are expected to drive popularity more than the worst keyword in an article. Refer to Figure 3.4.4 for the **confidence intervals** of each coefficient. The Deviance and AIC decrease from the model without interaction but that is a small decrease. One point to note here is



Logit(is_popular) = $\beta_0 + \beta_1$(num_keyword) + β_2*(kw_avg_avg) + β_3*(kw_best_avg) + β_4*(kw_worst_avg)*		
Coefficient – Estimate (odds-ratio scale)	p-value	Confidence interval (on Odds ratio scale)
beta$_0$ = 0.0024	<2e^{-16}	[0.0007,0.0021]
beta$_1$ = 1.12	4.72e-05	[1.06, 1.18]
beta$_2$ = 1.0002	< 2e-16	[1.0002,1.00025]
beta$_3$ = 1.000022	1.77e-08	[1.00001,1.00003]
beta$_4$ = 0.9998	0.00308	[0.9997,0.9999]



```
Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -7.455e+00  3.403e-01 -21.906  < 2e-16 ***
kw_worst_avg                  2.706e-05  3.364e-04   0.080 0.935890
kw_best_avg                   4.239e-06  6.525e-07   6.497 8.18e-11 ***
kw_avg_avg                    4.577e-04  6.017e-05   7.607 2.80e-14 ***
num_keywords                  1.024e-01  2.799e-02   3.658 0.000254 ***
kw_worst_avg:kw_best_avg     -2.066e-10  8.554e-10  -0.242 0.809149
kw_worst_avg:kw_avg_avg      -1.675e-08  2.694e-08  -0.622 0.534034
kw_best_avg:kw_avg_avg       -5.609e-10  1.306e-10  -4.295 1.74e-05 ***
kw_worst_avg:kw_best_avg:kw_avg_avg  3.193e-14  5.868e-14   0.544 0.586311
```



```
                                           2.5 %      97.5 %
(Intercept)                         0.0002968932 0.001127092
kw_worst_avg                        0.9993679338 1.000686619
kw_best_avg                         1.0000029606 1.000005518
kw_avg_avg                          1.0003398219 1.000575777
num_keywords                        1.0486927931 1.170311809
kw_worst_avg:kw_best_avg            0.9999999981 1.000000001
kw_worst_avg:kw_avg_avg             0.9999999305 1.000000036
kw_best_avg:kw_avg_avg              0.9999999992 1.000000000
kw_worst_avg:kw_best_avg:kw_avg_avg 1.0000000000 1.000000000
```

Table 3.4.1(left): The point estimates, confidence interval (both on odds ratio scale), and p-values for model without interaction.
Figure 3.4.3 (top): Results of logistic regression with interaction term (log odds scale).
Figure 3.4.4 (bottom): the 95% confidence interval for each variable on the odds-ratio scale.

3. Likelihood Ratio Test: At the 0.05 significance level, the chi-squared test yielded significant results with a p-value of **0.001215** and deviance of **23.09** indicating that the interaction term is indeed statistically significant and should be used in the model.

Model	Deviance	AIC	BIC
Null model	4736.92	-	-
Model 1	4562.81	4572.83	4615.74
Model 2	4539.72	4557.71	4634.9

Table 3.4.2: Compares the Deviance, AIC and BIC between the model without interaction and model with interaction. Deviance and AIC decrease while BIC slightly increases.

3.4.4 Conclusion

To conclude, the results from the logistic regression and Likelihood ratio test above support the claim that using the four explanatory variables – *kw_best_avg*, *kw_worst_avg*, *kw_avg_avg* and *num_keyword* along with an interaction between the first three terms yield the best model which has a reduced residual deviance and AIC score. However, the BIC score, which penalizes higher for a greater number of parameters than AIC, increased when the interaction term was added. Though these values are statistically significant, their odds ratio is very close to one for all the coefficients. This means that for a unit increase in the average number of shares for several keywords, there is very little (almost negligible) change in the odds ratio. Further domain knowledge is required to properly assess and understand the practical significance of these results.

3.4.5 Limitations and future work

A few limitations to the above work are as follows. There is a huge class imbalance between the popular and unpopular categories. Though all groups have greater than 15 counts, we might get better results from a more balanced dataset. I think procedures such as oversampling and under-sampling can be investigated for this. Another limitation is that around 600 rows were removed because they had negative values for *kw_worst_avg* field. This could lead to a potential loss of information. Filling this with dummy data or setting the values to 0 might help. One more assumption made throughout this analysis is that the data that we have is a good representation of the population data. If this doesn't hold, the results we have achieved will not be usable in real life.

3.5 Q5: Do title length and content length impact the popularity of an article?

- Navya Eedula

3.5.1 Data

The data for this section considers the explanatory variables *n_tokens_title* and *n_tokens_content* and later in the section, the categorical variables of the different data channels. It is important to note that in the original dataset, a logarithmic transformation was performed to scale these unbounded numeric features. The response variable chosen was the logit probability of article popularity, computed based on the predefined threshold for the number of shares, as described in Section 2.1.

3.5.2 Checking for assumptions

The assumptions to perform a logistic regression have been satisfied. The following assumptions were checked for - *(a)* binary nature of response variable, *(b)* linearity in log odds of the outcome variable with respect to the predictor variable, *(c)* multicollinearity. The code and details regarding this can be found in 6.5.

3.5.3 Statistical Methods

3.5.3.1 Evaluating Model Considering Title Length and Content-Length

Both content and title lengths exhibit skewness towards positive values. Content length ranges from -1.16 to 16.83, with a median on the shorter side (-0.29) and an average of 0, though skewed. Title length ranges from -3.97 to 5.96, with a median also on the shorter side (-0.19) and an average of 0, similarly skewed. These summaries indicate that while there's a spread in both lengths, their distribution tends towards longer values, potentially affecting their representativeness.

A logistic regression model was employed to analyze the relationship between title length, content length, and article popularity. Data scaling was performed on these two explanatory variables such that they are scaled between the values 0 and 1 to enable direct comparison of the impact of the two variables and their respective change in the log odds on the outcome. Initially, the individual effects of title length and content length on the logit probability of popularity. Subsequently, the combined effect of both variables was explored and further, the interaction effects between title length and article categories was tested. The observations are recorded in detail in Table 3.5.1.

Model	Feature	Coefficient	Significance	AIC	Interpretation
logit(is_popular) = β_0 + β_1*[n_tokens_title]*	n_tokens_title	0.15846	***	4730	Longer titles are associated with increased popularity.
logit(is_popular) = β_0 + β_1*[n_tokens_content]*	n_tokens_content	-0.09358	-	4737.7	Longer content is not significantly associated with popularity.
logit(is_popular) = β_0 + β_1*[n_tokens_title] +* β_2*[n_tokens_content]*	n_tokens_title	0.16043	***	4728.6	Both title and content length are included, but only title length has a significant positive effect
	n_tokens_content	-0.09657	-		
logit(is_popular) = β_0 + β_1*[n_tokens_title]* + β_2 *[n_tokens_content]* + β_3 *[n_tokens_title]* [n_tokens_content]*	n_tokens_title	0.15864	***	4730.2	Neither main effect of content length nor their interaction is significant; title length has a significant positive effect
	n_tokens_content	-0.09029	-		
	n_tokens_title * n_tokens_content	-0.03330	-		

*Significance codes: "***" (p-value < 0.0), "- " (not significant).*
Lower AIC indicates a better model fit (considering both deviance and complexity).

Table 3.5.1 Summary of the Logistic Regression Models Involving the Explanatory Variables of *n_token_title* and *n_token_content*

In summary, we observe that the coefficient for title length is statistically significant and positive. This means that a one-unit increase in the title length is associated with a positive change of **0.16** in the log odds of an article being popular. The confidence interval for the coefficient of the explanatory variable is **[0.066, 0.25]**. Since this confidence interval doesn't include 0, it indicates a statistically significant relationship between title length and the response variable. The positive confidence interval confirms that a one-unit increase in the explanatory variable correlates with a rise in the log odds of popularity. Neither the content length nor the combined effect of the title length and content length seem to show a significant association with the response variable. This observation is further strengthened by the confidence interval of the coefficient of content length variable **[-0.21, 0.0058]**, as it contains 0.

3.5.3.2 Evaluating Model Considering Title Length across the Data Channels

To further the analysis, I aimed to investigate how the effect of title length on popularity might vary across the different data channels. To uncover potential variations in both the mean and spread of title length across different data channels, box plots were generated as shown in Figure 3.5.1.

Analyzing the data shows variations in average article popularity across different channels. “Other”, “Lifestyle”, and “Social Media” channels have higher average shares, indicating greater popularity. In contrast, the “World” channel has the lowest average shares, suggesting lower article popularity. Moreover, there is notable variability in shares, with “Other” and “Business” channels experiencing significant variability, while the “Social Media" channel demonstrates the least variability.

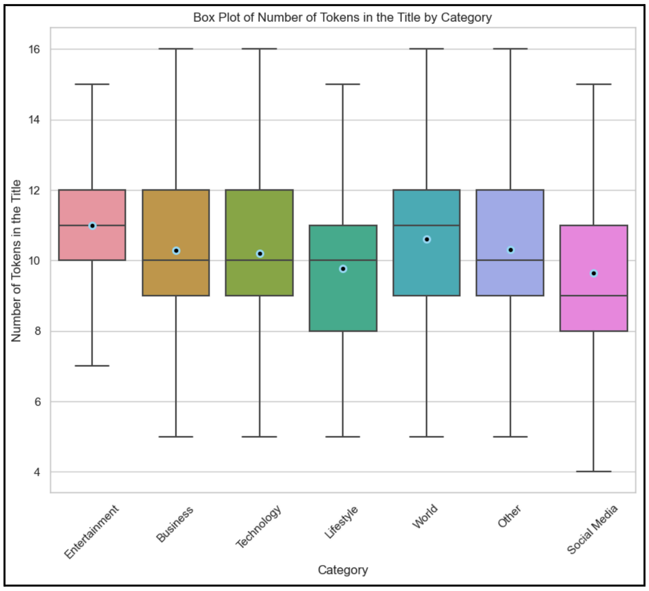


Figure 3.5.1 Distribution of Title Lengths across Categories

Firstly, a logistic regression model was employed to analyze the effect of the title length and the data channel on the response variable of popularity. Similar to the previous analysis, a longer title is associated with an increased chance of popularity because of a positive and significant coefficient (p-value = **0.00059**). The 95% confidence interval corresponding to the title length is **[0.073, 0.27]**, the range of which strengthens the evidence for the positive association. However, it is to be noticed that the value of the coefficient for title length is slightly larger than that of the previous analyses. However, it is to be noticed that the value of the coefficient for title length is slightly larger than that of the previous analyses. The strength of the apparent effect of the title seems to increase, and this could be due to providing similar information, making the effect of this variable seem stronger once it is accounted for.However, this could also mean that there may be a more complex scenario where the title length and category are both influenced by a third unobserved variable. The coefficients for each category represent the effect on popularity compared to the baseline (which is the Category: Business). Category "Other" is the most popular category (highest coefficient, **1.39**, and highly significant **p-value < 2e-16** with a 95% confidence interval of **[1.083,1.72]**).

When the interaction between the title length and each of the categories was added, this seemed to yield very interesting results. Similar to the previous case, most data channels show no significant effects. The title length is still positive and significant (with a p-value of **1.46e-09** and a 95% confidence interval **[0.0053, 0.0104]** to support the same). However, we observe a large decrease in the magnitude of the coefficient which may imply that incorporating new explanatory variables shares some explanatory power with title length. The original effect of title length may still exist but is now partly obscured by the inclusion of the interaction of different data channels. In the previous case, the data channel "Other" seemed to be fairly positive and highly significant. However, this effect vanished in the interaction model, possibly due to the weakening of individual coefficient effects when the interaction model was introduced. It's also possible that the non-interaction model was overfitting, and the more complex model with interaction was better able to accommodate it. Interestingly, the impact of title length varied across different content categories, with "Technology" and "Entertainment" articles exhibiting distinct patterns compared to other categories. "Entertainment" articles with longer titles tend to be less popular than the "Business" category (negative and marginally significant coefficient with a p-value of 0.038 and 95% confidence interval ranging from **[-0.00039, 0.0068]**). Similarly the data channel, "Technology" shows a stronger negative association with popularity for longer titles compared to "Business" (negative coefficient for the main effect of "Technology" and significant negative interaction term with p-value = 0.0066 and 95% confidence interval ranging from **[-0.0082, -0.0013]**). Most importantly, the AIC for the interaction model is much lower (**-67634**) compared to the previous model (which was **4523.8**), suggesting a potentially better fit for the data.

Although the model showing interaction demonstrated interesting results, it is not a very simplistic model that can be easily interpreted. It requires further analysis of the interaction terms to understand how category and title length influence popularity together.

4 Discussion and conclusion

Our analysis set out to investigate what are the different factors that affect the popularity of a news article.

In question 1, the chi-square test showed that the proportion of articles containing multimedia content (images/videos) is not the same across categories. The coefficients of logistic regression models with *num_imgs_std* and *num_videos_std* indicated a positive and statistically significant influence on its popularity. When *category* was incorporated into the model, the effect of video (*num_videos_std*) became statistically insignificant. Notably, articles belonging to the "Other" category exhibited a slightly higher likelihood of popularity compared to those in the "Business" category, after adjusting for image and video content. These findings suggest that both images and the category of an article contribute to its popularity. However, further investigation is needed to determine the practical magnitude of these influences.

In question 2, the study conclusively shows that the data channel of online articles significantly affects their popularity, as measured by shares which resulted from ANOVA analysis and Tukey HSD test for pairwise comparison. This analysis underscores the importance of content strategy alignment with audience preferences and the exploration of underrepresented categories for enhancing article visibility and engagement. Particularly, articles within the "Other," "Lifestyle," and "Social Media" channels tend to achieve higher popularity, highlighting the varied interests of online audiences. Conversely, the "World" channel's articles are less popular, indicating potential discrepancies in audience engagement or content delivery. Further investigation into the "Other" category could reveal new insights for content optimization.

In question 3, our analysis explored how publication day and channel impact news article popularity. We found weekends consistently generated higher shares than weekdays, with Saturday articles outperforming those released on Tuesday and Thursday. While weekdays saw more articles published, weekends were preferred for garnering shares. Variation in popularity across channels, especially on weekends, was observed. Notably, the "world" channel had lower performance than others, except business, particularly on Sunday. However, Bonferroni's correction's conservatism may have hindered identifying superior channels on Saturday. These results underscore the complex interplay of publication day and channel on article popularity, warranting further exploration to refine statistical approaches and understand underlying mechanisms.

In question 4, we concluded that keywords in an article statistically impact the log odds of an article being popular. After fitting logistic regression with and without interaction and performing a Likelihood ratio test for the significance of the interaction, the model with the four explanatory variables – kw_best_avg, kw_worst_avg, kw_avg_avg, and num_keyword, along with interaction among the first three, results in the most optimal model, as evidenced by reduced residual deviance and AIC score. However, these results must be used cautiously as the narrow confidence intervals of the log odds which are very close to 1 indicate that the model while being statistically significant might not have much practical significance.

In question 5, the findings underscore the importance of title length in determining article popularity. Furthermore, it was observed that the effect of title length on the popularity of the article was different for the data channels ``Entertainment” and “Technology” compared to other data channels. This may suggest that content creators should carefully consider title length when crafting headlines. Moreover, the observed differences in the effect of title length across article categories emphasize the need for tailored strategies in content marketing and promotion.

In our current analysis, we have defined the top 1% of articles as "popular." This high imbalance in data may lead to potential issues. When the majority class dominates it leads to biased coefficients favoring the majority class predictions. Additionally, accurately estimating coefficients for minority class variables is challenging due to limited observations, potentially resulting in unreliable interpretations of coefficients and misleading conclusions. One approach that can be used here is to oversample the minority class or undersample the majority class. However, more research is required to determine the exact methodology to be used.

5 References

[1] Fernandes, Kelwin et al. “A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News.” Portuguese Conference on Artificial Intelligence (2015).

[2] Fernandes,Kelwin, Vinagre,Pedro, Cortez,Paulo, and Sernadela,Pedro. (2015). Online News Popularity. UCI Machine Learning Repository. https://doi.org/10.24432/C5NS3V.

6 Code

Our code is located in this GitHub Repository